Future Storage Solutions for Data Analysis

ACAT 2013

Robert Triendl

Vice-President, DDN Japan

Topics

- CPU vs. Storage
  - Balance between CPU performance and storage performance
  - Opportunities and challenges
- Storage Devices
  - Flash/NVRAM vs. Disks
  - Novel storage hierarchies
- Object Storage
  - Storage for applications (not humans)
  - POSIX vs. Object APIs

Storage for HPC and Data Analysis: The Past Decade

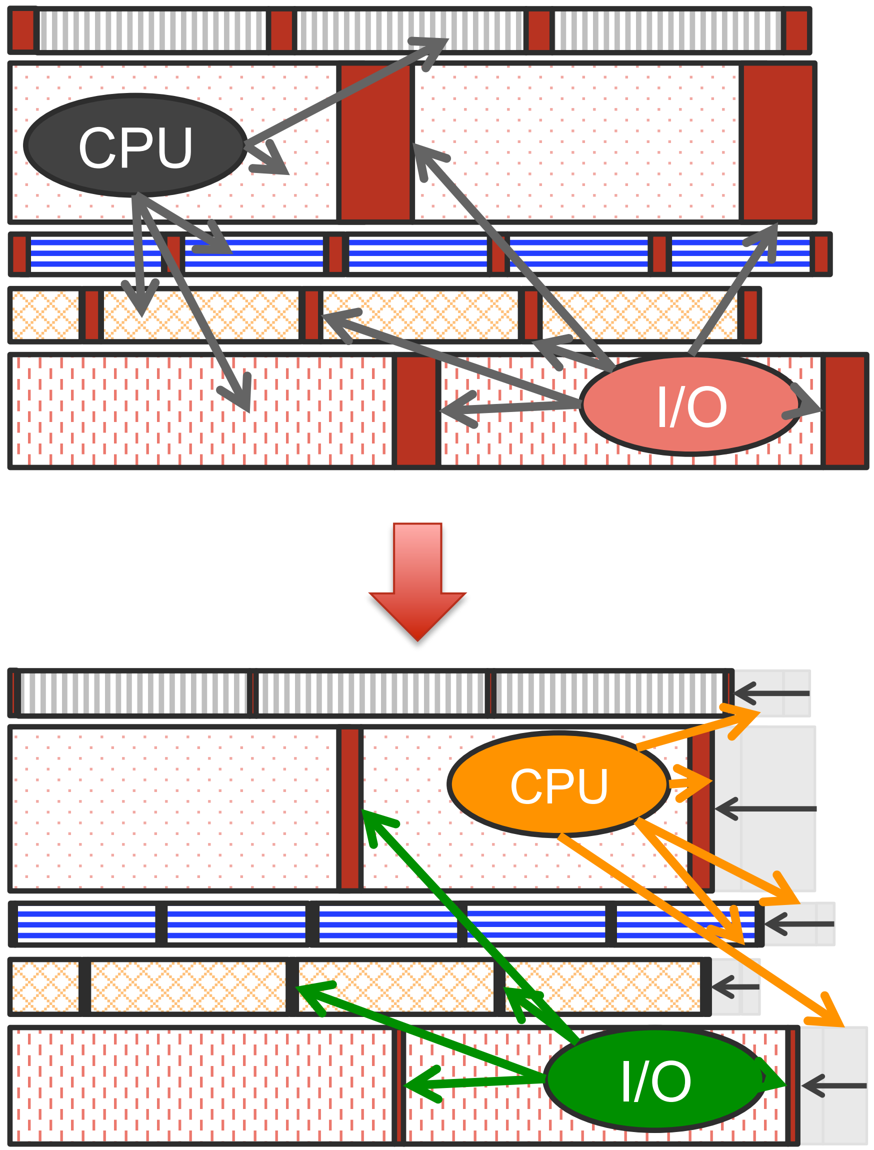
- **I/O as Problem for Computation**
  - Dealing with PBs of data and billions of files
  - I/O as a performance bottleneck for computation
- **Emergence of “I/O Clusters”**
  - Shared storage for different compute clusters
  - Wide-area access to data
- **Parallel and Distributed I/O**
  - Parallel file systems: Initially many flavors, now mainly Lustre, General Parallel File System (GPFS), and some flavors of parallel NFS (pNFS)
  - Distributed I/O: Hadoop/HDFS, Gluster, Ceph, G-farm, various flavors of commercial object-storage
  - Open-source, “software-defined storage” vs. expensive solutions from storage vendors

I/O as a Problem

- Disk Performance
  - 70-150 MB/sec per SATA/NL-SAS device
  - Access latency has barely changed over the past two decade
- Device Interfaces
  - Disks today are 6 Gbps SAS
  - 12 Gbps SAS is coming
  - Will provide above 3 GB/sec per interface…
- New Semiconductor-Based Devices
  - Still expensive, but cost is coming down quickly!
  - Next generation SSDs will have capacities of a few TB!
  - Up to GBs/sec for Flash/NV-RAM devices

Why Storage Matters

- Computer vs. Storage Cost
  - Typically 10-15% of the cost for an HPC system is spent on I/O and storage
  - The ration can be significantly higher for data analysis systems (up to 35 % spent on storage, but very rarely more)
- Optimal Compute/Storage Investment Ratio
  - Not simply storage capacity and peak performance!
  - But, rather, optimal ration depends critically on the time needed for I/O vs. time needed for compute
- I/O Intensive Applications
  - Not simply the amount of data transferred between nodes and storage
  - Applications can be either transactional (IOPS) or streaming (sequential I/O), depending on the way how the application actually reads and writes data from storage…

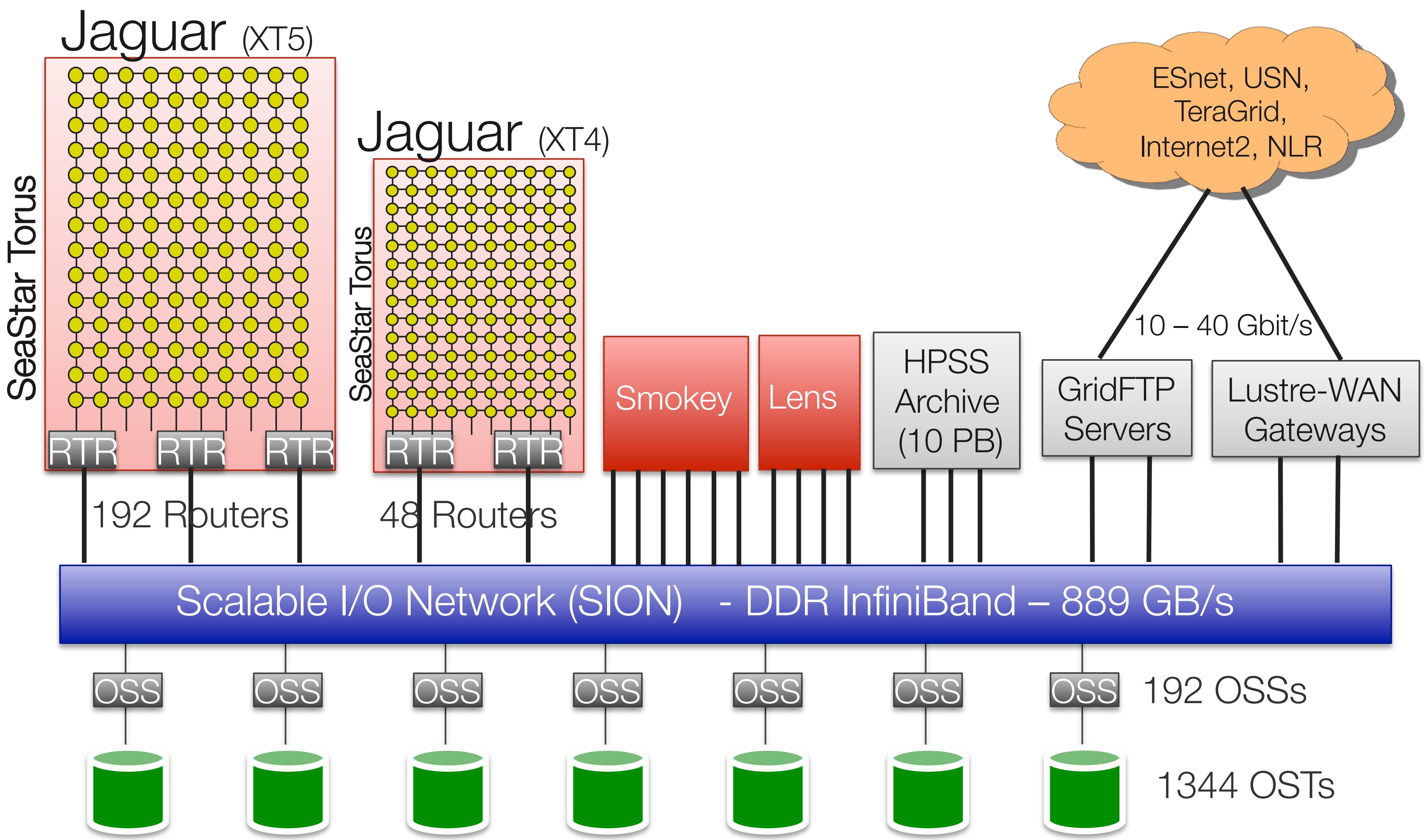


ddn.com

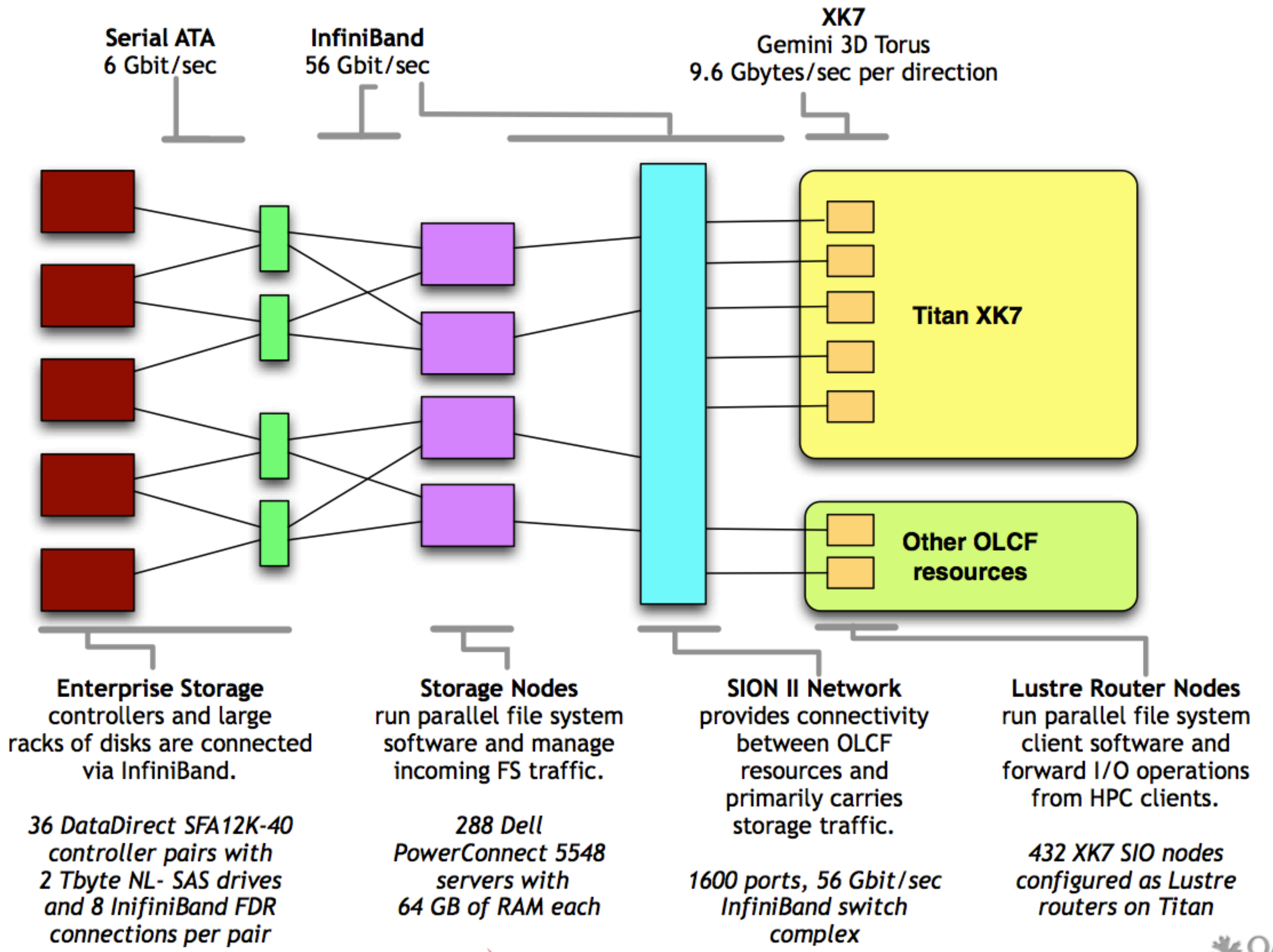
I/O Clusters

- I/O & Storage
  - Fast I/O used to “local” to a given compute system
  - Shared tape (or “nearline” disk) archive as backend
- Issues
  - Staging of I/O is needed
  - Various performance problems
- Today
  - Many compute clusters accessing the same global storage environment over 10/40 GbE or IB storage network
  - Option to access storage over WAN is becoming available
  - Linkage from fast storage to grids/clouds

I/O Clusters: Example Oak Ridge (2008)



Oak Ridge Spider II File System



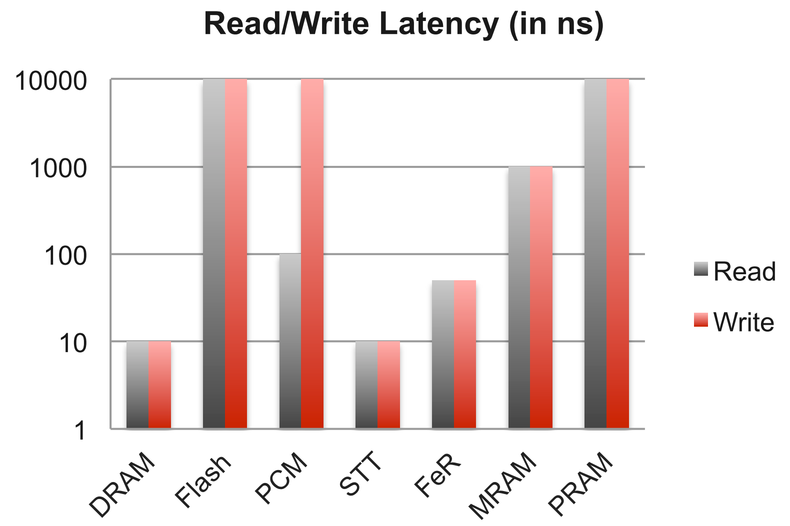
SPIDER II

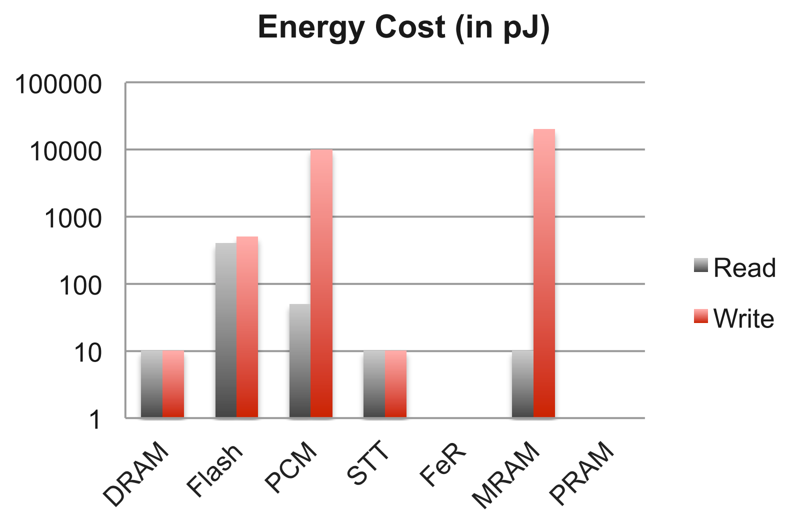
Parallel File Systems

- Used to be "Exotic"
  - Difficult to install and administrate
  - Full of bugs
  - Very poor metadata performance (due to distributed locking etc.)
  - Limited RAS features (leading to downtime, data loss, etc.)
  - Limited usability (Linux kernel limitations)
- Very Common Today
  - Parallel file systems are very common in both HPC and data analysis
  - Stability has improved significantly, even with open source file systems
  - Metadata performance has improved significantly
  - Depending on requirements, not that many options left…
  - Shift toward distributed architectures?

Bringing Data to the CPU

- **Device Options**
  - SRAM, DRAM: small and fast memory
  - FLASH, PCM: higher capacity but less I/O activity
- **Rethinking I/O Hierarchies**
  - Memory Bus: Memory | PCM
  - I/O Bus: FLASH
  - Cluster Fabric: FLASH
  - SAN: Disks | FLASH





Data from Mark Seager, Intel.

DDN SFX and DDN Burst Buffer

<u>DDN Burst Buffer (“Global Cache”)</u>

Flash / NVRAM cache for file system accesses

Wedges between HPC applications and file systems

Distributed cache for file system namespaces

BW optimized

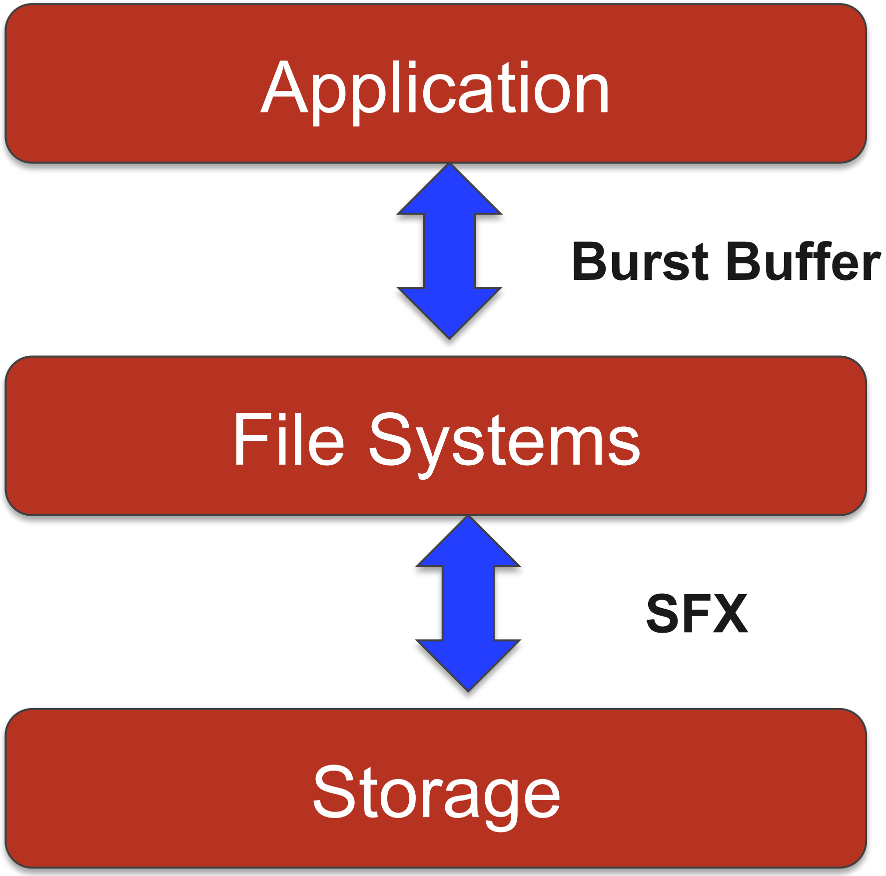
<u>DDN SFX</u>

Flash cache for block device accesses

Integrates with SFA OS (enterprise storage block device)

Per-block device cache

IOPs optimized



Iron Monkey – Overview

- Flexible burst buffer implementation
  - Supports various degrees of fault tolerance
  - Supports various deployment modes
- Targeted at both extreme scale and mid-range commercial HPC
- Removes PFS from the I/O path for bulk data accesses
- Isolate and / or optimize ill-behaving applications with sub-optimal I/O patterns

Hyperscale Storage: HPC and Cloud

High Performance Computing	Cloud Computing
Mostly very large files (GBs)	Small and medium size files (MBs)
Mostly write I/O performance	Mostly read I/O performance
Mostly streaming performance	Mostly transactional performance
10s of Petabytes of Data	Billions of files
Scratch data	WORM (Write-Once-Read-Many)
100,000s cores	Millions of cores
Mostly Infiniband	Mostly Ethernet
Single location	Highly distributed data
Limited replication factor	High replication factor

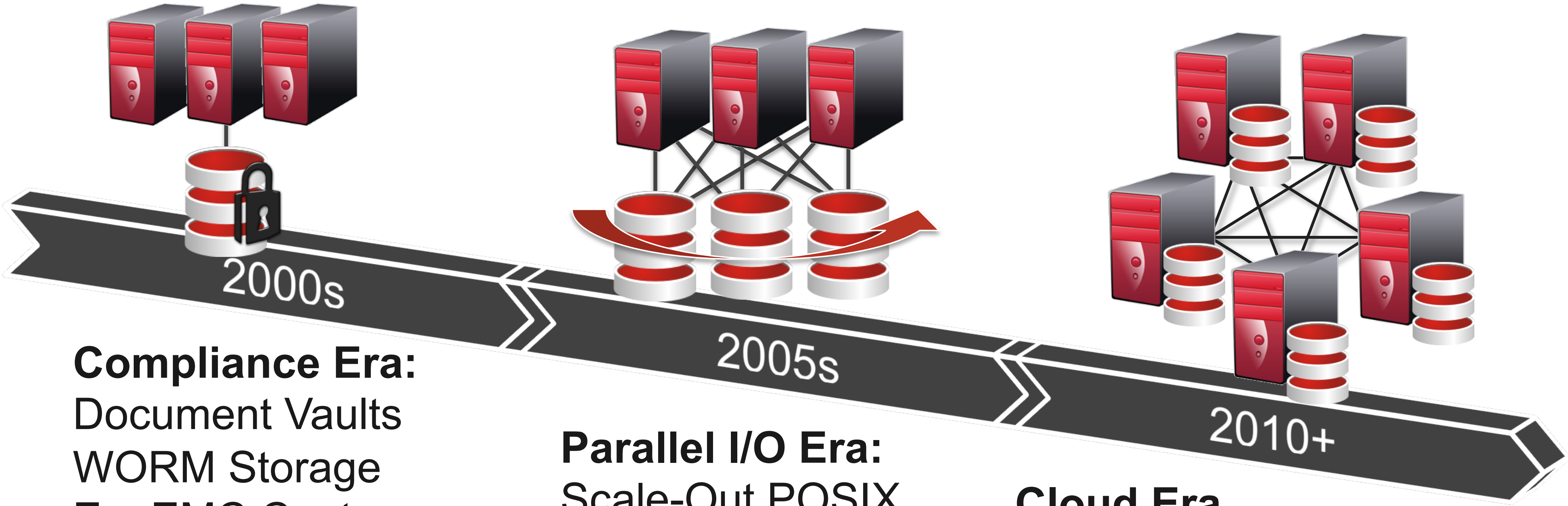
Data Analysis

Example Amazon S3

- Amazon S3 (**S**imple **S**torage **S**ervice)
- Object storage cloud launched in 2006
- Amazon S3 API
- Presently stores well over a trillion of objects, organized in “buckets” (owned by an AWS account)
- REST or SOAP Interface – can be accessed by unmodified HTTP clients, so easy to replace existing web hosting infrastructures
- HTTP Get or BitTorrent protocols
- Users include DropBox, Zmanda, StoreGrid, Minecraft, etc.

Object Storage

Challenges & Opportunities



Compliance Era:
Document Vaults
WORM Storage
Eg. EMC Centera

Parallel I/O Era:
Scale-Out POSIX
Eg. Lustre, GPFS
(not discussed as object storage)

Cloud Era
Peer:Peer, Hyperscale
Eg. Amazon S3

- **Object storage's history in the archive and compliance market has created an impression in the market that object storage is for archive only.**
- **POSIX-applications are difficult to integrate with object storage interfaces.**

Object Storage

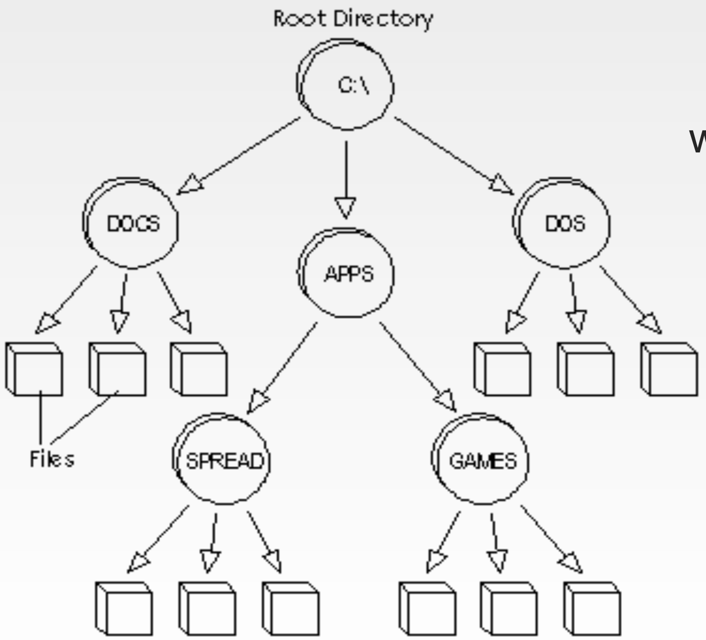
Storage for Humans

`User/Data/Powerpoint/WOS`

Storage for Applications

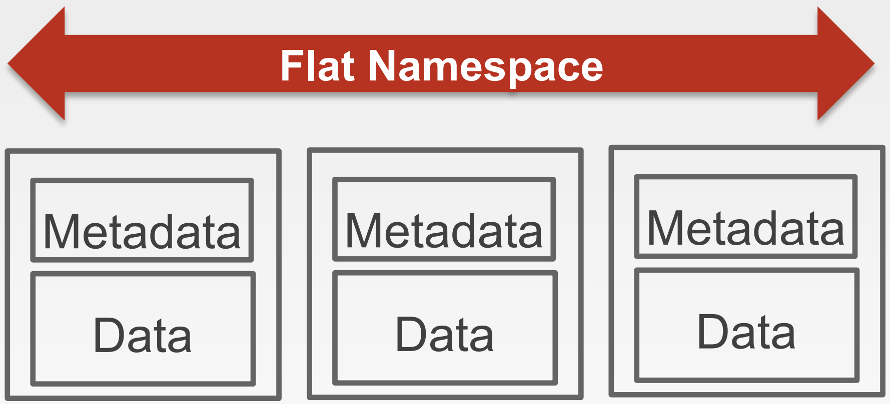
`ACuoBKmWW3Uw1W2TmVYthA`

File Systems



File Systems were designed to run individual computers, then limited shared concurrent access, not to store billions of files globally

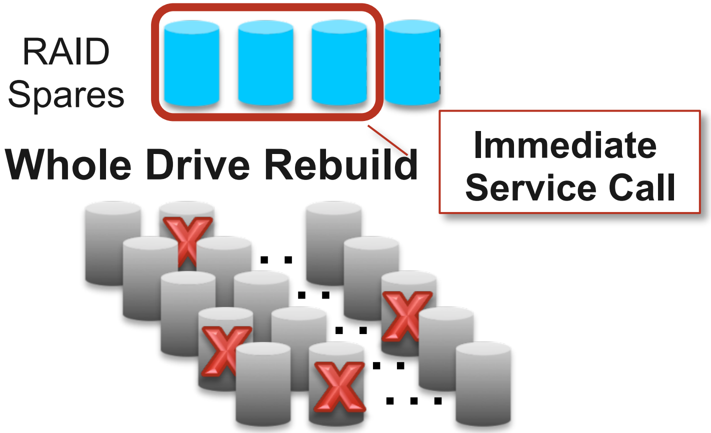
Objects



Objects are stored in an infinitely large flat address space that can contain billions of files without file system complexity

Intelligent Data Protection with OA
RAID vs WOS DeClustered Re-Balance

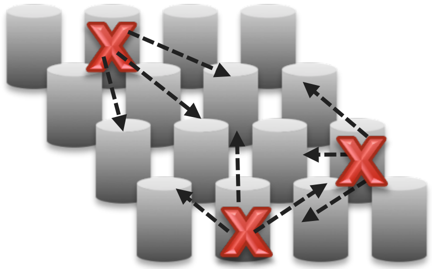
FS + RAID 6



Traditional RAID Storage

WOS

Re-Balance



Capacity Available: 120TB

Optional Scheduled Service Call Restores Capacity

WOS (Replicated or Object Assure)

RAID Rebuilds Drives

Lost capacity - Spare drives strand capacity

Long rebuild times - Whole drive must be rebuilt even though failed drive only partially full

Higher risk of data loss – if spare drive is not available, no rebuild can occur

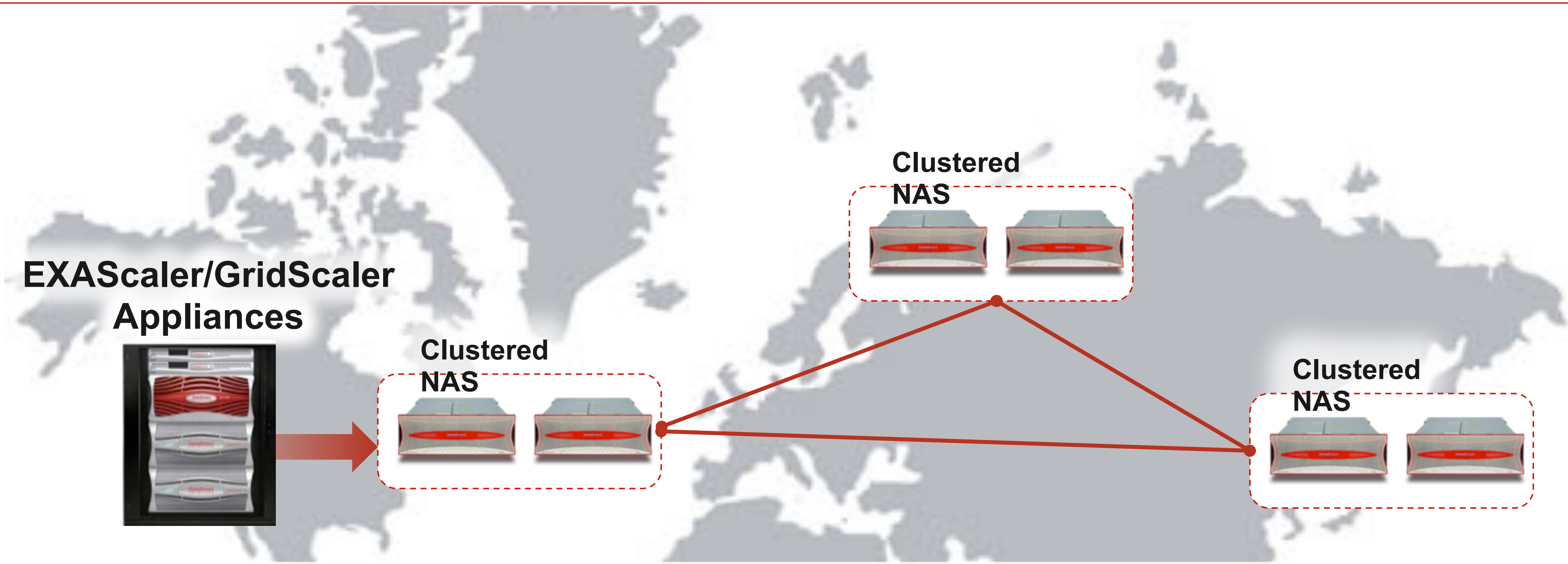
Increased support costs - immediate service call is required to replace low spares condition

Reduced write performance- RAID reduces disk write performance, especially for small files

WOS Re-Balances Data Across Drives

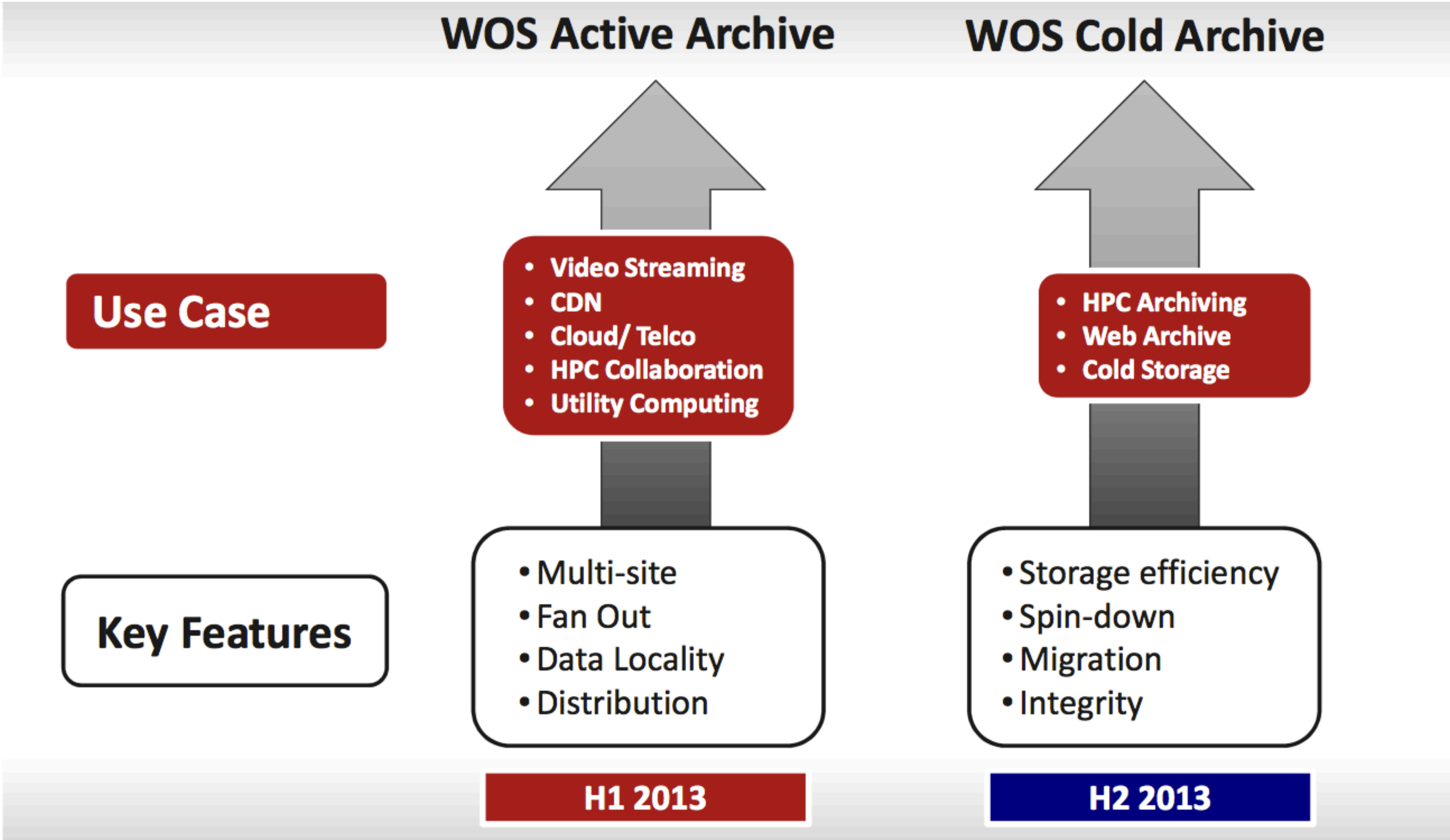
- All drives fully utilized – Any free capacity on any drive is part of the spare pool
- 50%+ shorter re-balance times – Only actual data is copied, rebuild at *read* speeds, not *write*
- Faster recovery times increase overall performance and reduce risk of data loss
- Drive failures decrease overall capacity only by the size of the failed drives
- Total capacity may be restored by replacing drives during scheduled maintenance

Automated, Cloud-Based Collaboration



- **Cloud Ready -** Tiering ready with file system tiering to a public or private WOS cloud - share and disseminate information globally
- **Collaboration Ready -** Eliminate organizational storage silos while automating data distribution & collaboration
- **Archive Ready -** Backup files safely to a public or private WOS cloud for disaster recovery

WOS Future


WOS Active Archive
WOS Cold Archive
Use Case
• Video Streaming
• CDN
• Cloud/ Telco
• HPC Collaboration
• Utility Computing
• HPC Archiving
• Web Archive
• Cold Storage
Key Features
• Multi-site
• Fan Out
• Data Locality
• Distribution
• Storage efficiency
• Spin-down
• Migration
• Integrity
H1 2013
H2 2013

DataDirect™
NETWORKS
INFORMATION IN MOTION™

Thank you!